

# Preface

*Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism* is an anthology of the research findings of thirty-five speaker recognition experts from around the world. The book provides a multidimensional look at the complex science involved in determining whether a suspect's voice truly matches forensic speech samples, collected by law enforcement and counter-terrorism agencies, that are associated with the commission of a terrorist act or other crime. Given the serious consequences for the suspect, who may go to jail or even (in the most extreme cases involving terrorism or murder) face the death penalty, a rigorous and reliable science must be used in finding a match between a suspect's voice and the speech samples collected in the forensic crime lab. The United States Supreme Court's ruling in *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) established a test for the legal admissibility of scientific evidence which requires that the theory and method upon which the evidence is based is testable, accepted, peer reviewed, and, where applicable, has a known equal error rate (EER). Similar standards for the validity and reliability of scientific evidence are used in other countries that have taken up the recommendations of the *National Academy of Sciences* (USA) and the *Law Commission* (UK).

Standards like these place a heavy burden on the expert who offers testimony in court. Each time forensic testimony is entered into evidence, the expert witness must prove *ab initio* that his science is reliable and valid in the case that is before the court before his or her testimony can properly qualify for admissibility. It follows that if forensic scientific methods are to be useful in legal contexts, they must hold up under judicial scrutiny, since in the end the question of admissibility will be decided by a trial judge.

Hence the consistent stress on bringing speaker authentication methods into line with the strict standards of legal admissibility is exactly what the reader will find in the work of this volume's diverse group of forensic speech scientists, whether they work side by side with investigators in crime labs, provide services to private companies that specialize in the design of speaker verification systems, or teach in university settings where they study (among other things) the effects of speech signal degradation on the quality of forensic speech samples. Forensic speaker recognition, as a probative science, must competently assist criminal investigators in

minimizing both the occurrence of a “false positive”—in which the speech sample related to the commission of a crime or terrorist act is matched to the wrong suspect—or a “false negative,” in which the real culprit’s voice fails to match the crime lab’s speech sample meant to fit him.

Although divided into eighteen chapters, addressing such varied topics as the challenges of forensic case work, handling speech signal degradation, analyzing features of speaker recognition to optimize voice verification system performance, and designing voice applications that meet the practical needs of law enforcement and counter-terrorism agencies, this book’s material all sounds a common theme: how the rigors of forensic utility are demanding new levels of excellence in all aspects of speaker recognition. The book’s contributors are among the most eminent scientists in speech engineering and signal processing; their work represents the best to be found at universities, research institutes for police science, law enforcement agencies and speech companies, in such diverse countries as Switzerland, Sweden, Italy, France, Japan, India and the United States.

*Forensic Speaker Recognition* opens with an historical and procedural overview of forensic speaker recognition as a science. Following this is a fascinating exposition by Professor Andrzej Drygajlo of the Swiss Federal Institute of Technology in Lausanne, whose chapter focuses on “the research advances in forensic automatic speaker recognition (FASR), including data-driven tools and related methodology that provide a coherent way of quantifying and presenting recorded voice as biometric evidence.” Professor Drygajlo furnishes the reader with an in-depth discussion of the

European Network of Forensic Science Institute’s evaluation campaign through a fake (simulated) case, organized by the Netherlands Forensic Institute, as an example where an automatic method using the Gaussian mixture models (GMMs) and the Bayesian interpretation (BI) framework were implemented for the forensic speaker recognition task.

This first section, aptly titled “Forensic Case Work,” is further enriched by the investigations of Swedish professor Anders Eriksson (of the University of Gothenburg) into the specific challenges of forensic case work. Drawing on a substantial number of investigations performed for the Swedish police, the author inspects in painstaking detail the differences between the aural/acoustic and the automatic methods in forensic case work, focusing on what works and what doesn’t in real-life settings. The section concludes with a fascinating study of speaker profiling, based on the characteristics associated with speaker dialect. Manisha Kulshretha, a Haskins’ Laboratory (Yale University) researcher, together with C. P. Singh of the Forensics Science Laboratory, Government of NCT of Delhi, and Professor R. M. Sharma of Punjab University, show that from a sample size of 210 speakers, acoustic features associated with lexical tone and sentence intonation, along with vowel quality and vowel duration, serve potentially to identify the speaker’s particular dialect. Where dialect is an important element of identification, this method helps investigators to appreciably narrow the pool of potential suspects to those who reside in the region where that particular dialect is spoken.

The second section of the book, titled “Speech Signal Degradation: Managing Problematic Conditions Affecting Probative Speech Samples,” devotes considerable attention to the stubborn problem of speech signal degradation that impedes the gathering of probative speech samples (that is, samples gathered for use in court processes) that are clear and audible. Since criminals, in their zeal to cover their tracks, often lower their voices even to a whisper, or make calls from public places where there is loud noise in the background (or use VoIP networks) the quality of the voice recording is often poor. Thus, much of the speech data available for forensic analysis are degraded by several factors such as background noise, transmission and channel impairments, microphone variability, multi-party conversations, whispered speech, and VoIP artifacts. As a result, speech scientists, as part of their efforts to manage the problematic conditions affecting the quality of probative speech samples, must carefully isolate and measure the effects of all such factors on speech signal degradation.

The authors presented in this section have met that challenge head-on. They bring to the discussion the results of years of careful study of degraded speech on the performance of an automatic speaker recognition (ASR) system, by concentrating on the following problems and, where available, their possible solutions:

1. speech under stress and the “Lombard Effect”;
2. the wide range of artifacts of VoIP (speech codec, packet loss, packet reordering, network jitter, foreign-cross talk or echo) and the effect of such artifacts on the performance of an ASR system;
3. session variability (“mismatched” environments for collection of speech samples) and the use of the non-linear modeling techniques of Teager Energy Operator-based Cepstral Coefficients (TEOCC) and amplitude versus frequency modulation (AM-FM) to improve speaker recognition in mismatched environments;
4. noisy environments and the use of speaker-specific prosodic features to improve speaker recognition;
5. noisy backgrounds and the use of various noise reduction filters (Noise Reduction, Noise Gate, Notch Filter, Bandpass, and Butterworth Filter) in enhancing the speech signal for speaker identification; and
6. whispered speech and the use of an algorithm for whisper speech detection as part of a seamless neutral/whisper mismatched closed-set speaker recognition system.

This section has far too many contributors to name each one individually. They include University of Texas Professor John H. L. Hansen, University of Minnesota Professor Keshab K. Parhi, Raghunath S. Holambe, professor at SGGS Institute of Engineering and Technology, Nanded, India, and Jiju P. V., Senior Scientific Officer at the Forensic Science Laboratory, Government of NCT of Delhi, among other distinguished speech signal experts.

The third section, titled “Methods and Strategies: Analyzing Features of Speaker Recognition to Optimize Voice Verification System Performance in Legal Settings,” presents the experimental research findings of some of the most innovative and forward-looking speech scientists who have isolated the important features of speaker

recognition (some of which are appreciably less affected by signal degradation than others), and have carefully analyzed how such features may play an important role in improving forensic automatic speaker recognition.

The section begins with the experimental findings of Kanae Amino of the National Research Institute of Police Science in Japan (together with her research collaborators), showing that nasal sounds are effective for forensic speaker recognition despite the differences in speaker sets and recording channels. They show how “performance degradation caused by the channel difference, in this study of air- and bone-conduction ... can be redressed by devising normalisation methods and acoustic parameters.”

Next, T. V. Ananthapadmanabha, CEO of Voice and Speech Systems in Bangalore, describes his careful studies of the volume-velocity airflow through the glottis (or the glottal airflow). In so doing, he has explored the significance of speech source characteristics by utilizing rigorous analytical results from the aerodynamic and acoustic theory of voice production. Much of this work was inspired by the author’s research collaboration with the late Professor Gunnar Fant at the Royal Institute of Technology, Stockholm in the early 1980s. “A good understanding of the theory guides one in appropriate modeling and interpretation of voice source,” he writes. In addition, Dr. Ananthapadmanabha contends that “habitually formed relative dynamic variations in voice source parameters are of greater significance in forensic speaker recognition.”

The section is further enhanced by the analytic insights of Leena Mary, professor at Rajiv Gandhi Institute of Technology, Kottayam, India on the effectiveness of syllable-based prosodic features for speaker recognition. In her chapter, Professor Mary describes in painstaking detail a method for extracting prosodic features directly from the speech signal itself. “Applying this method,” she tells us, speech is segmented into syllable-like regions using vowel onset points (VOP). The locations of VOPs (which entail Hilbert envelope of the linear prediction (LP) residual signal) serve as reference for extraction and representation of prosodic features.”

Significantly, Professor Mary deliberately chose to analyze prosody—which reflects the learned/acquired speaking habits of a person and therefore contributes to speaker recognition—in as much as prosodic features are less affected by channel mismatch and noise, which are common causes of speech signal degradation in probative speech samples. Thus, prosodic features are particularly well suited to speaker forensics, a field that demands accurate identification of suspects and therefore a minimum of obstacles to robust speaker recognition, such as those posed by channel transmission problems.

The section is rounded off by the study findings of C. Chandra Sekhar, professor at the Indian Institute of Technology (IIT), Chennai, India, and his graduate student assistant, A. D. Dileep. The authors meticulously show that when the performance of Intermediate Matching Kernel (IMK)-based Support Vector Machines (SVMs) is compared to that of state-of-the-art GMM-based approaches to speaker identification (using the 2002 and 2003 NIST speaker recognition corpora in evaluation of different approaches to speaker identification), the IMK-based SVMs performed significantly better than the GMM-based approaches for speaker identification

tasks. From this comparison, the authors draw the conclusion that because IMK-based SVMs are well suited to the basic challenges of providing reliable scores for intra-speaker variation of suspects and for inter-speaker variation within a potential population, they can play an important role in serving the needs of law enforcement and counter-terrorism agencies in performing forensic speaker recognition.

The final section of the book, titled “Applications to Law Enforcement and Counter-Terrorism,” enlightens the reader about practical constraints in the use of forensic speaker recognition systems for the daily concerns of law enforcement and counter-terrorism agencies. The section begins with the research of V. Ramasubramanian, who serves as a senior member of Siemen’s (Bangalore) technical staff, on automated telephony surveillance to detect if a person from a specific government watch-list is on the line at a given moment. As he points out:

[S]uch an automatic solution is of considerable interest in the context of homeland security, where a potentially large number of wire tapped conversations may have to be processed in parallel, in different deployment scenarios and demographic conditions, and with typically large watch-lists, all of which make manual lawful interception unmanageable, tedious and perhaps even impossible.

His chapter begins with the “basic framework for watch-list based speaker-spotting, namely, open-set speaker identification, subsequently refined into a ‘multi-target detection’ framework.” Dr. Ramasubramanian examines in detail “the main theoretical analysis available within the framework of multi-target identification, leading to performance predictions of such systems with respect to the watch-list size as the critical factor.”

Taking an applications-oriented approach to forensic speaker recognition, he then outlines related speech topics—speaker change detection, speaker segmentation and speaker diarization—that can be useful in the design of automated telephony surveillance for border security and protecting critical infrastructure. These and other issues and concerns inhabit the broader context of homeland security. The author concludes with a summary of product level solutions currently available in the context of surveillance and homeland security applications, while acknowledging the realistic challenges and limitations faced by automated speaker-spotting systems.

Next, Patrick Perrot of the Forensic Research Institute of the French Gendarmerie and Gerard Chollet of Telecom-Paris take up the fascinating topic of criminals who disguise their voices to hide their actual identity, sometimes even impersonating someone else. Such disguises typically occur when criminals make telephone threats, malicious calls, extortion attempts and/or blackmail, or terrorist demands. The authors point out that while

there are those cases when there are involuntary voice changes, as when there are alterations in voice characteristics due to poor transmission of telephonic communication ... or even pathologies (both acute and chronic) that morph speech production ... we limit this discussion to disguise which consists of a person who *deliberately* conceals his identity ... as a means of misleading the human ear or even the automatic speaker recognition system.

Drs. Perrot and Chollet focus on specific voice characteristics to evaluate the recognition of a suspect's voice in the presence of voice disguise. Their analyses of voice transformation are based both on an acoustic approach, which they use to measure specific changes in speech, and on an automatic approach, which is employed to detect voice disguise. The acoustic analysis of specific features reveals that the effect of the disguise on voice characteristics is dependent upon the kind of disguise that is used, while in the automatic experiment the authors performed, they found that parallel fusion and SVM classifier provided the best results with a good level of discrimination.

The practical applications of these two French scientists' work can be seen in the fact that a major part of their research into voice disguise has been devoted to the study of voice disguise reversibility. Their studies of voice disguise reversibility have revealed that

while it is not possible today to fully reverse a voice disguise in such a way that the resulting waveform would sound completely natural to a listener (mainly due to limitations with the quality of converted voice synthesis), our study demonstrates, nevertheless, that a disguised voice could be reversed to a relatively "normal" voice as evaluated by current state of the art speaker verification systems.

Thus, the authors see a more robust speaker recognition on a reverse-disguised voice—that is, a voice that has already been converted back from its disguised form to normal speech—as a future practical application of their research, as well as evaluation of the performance of speech applications in such contexts.

The last two chapters of the book are authored by speech experts at Nuance Communications and Loquendo. Chuck Buffum, Nuance's Vice President, provides insightful lessons learned from commercial voice biometric deployments to forensic applications, giving the reader a better understanding of the evolution of speaker verification systems in forensic settings. Mr. Buffum points out that "commercial deployments of voice biometrics have predictably focused primarily on automating the correct acceptance of true users for telephony self-service. However, over the past few years, a trend has developed within the financial institutions to begin using voice biometric technology to look for duplicate enrollments or to investigate suspicious transaction activity," a trend that, he contends, "opens the discussion of bringing relevant techniques and experiences from commercial voice biometric deployments into the forensic voice biometric space."

Avery Glasser, consulting architect for the Italian-based company Loquendo, closely examines the practical needs of anyone wishing to implement investigatory voice biometric technology, and how best to bridge the gap between creators and implementers of this technology. The author points out that "there are critical problems that only voice biometrics can solve, but getting the solutions well positioned requires a deep understanding of the nature of government implementations that seems to escape the grasp of too many vendors. The chapter," according to Mr. Glasser's exordium, "will explore a number of critical use cases and provide perspective on how technology creators can position their solutions to meet those needs."

As the editors of this compendium we have endeavored to bring together notable forensic speaker recognition experts who, by virtue of their meticulous research and keen attentiveness to the needs of law enforcement and counter-terrorism agencies, have both individually and collectively brought forensic automatic speaker recognition (FASR) technology to a new plane. It is our hope that this science will continue to evolve so that the admissibility of speaker recognition evidence will no longer present a Sisyphean challenge to prosecutors who have come to depend on voice verification systems to make a convincing case to the court about the identity of a criminal suspect.

Fort Lee, NJ, USA  
Gandhinagar, India

Amy Neustein, Ph.D.  
Hemant A. Patil, Ph.D.



<http://www.springer.com/978-1-4614-0262-6>

Forensic Speaker Recognition  
Law Enforcement and Counter-Terrorism  
Neustein, A.; Patil, H.A. (Eds.)  
2012, XXI, 540 p., Hardcover  
ISBN: 978-1-4614-0262-6